probe amino acid sequence through different template structures and attempts to find the most compatible structure for a given sequence. In certain embodiments, sequence-to-structure alignments are performed by a "local-global" version of the Smith-Waterman dynamic programming algorithm (Waterman, 1995). In such embodiments, alignments are ranked by one or more, preferably three, different scoring methods. In a three method approach (Jaroszewski *et al.*, 1997), the first scoring method can be based on a sequence-sequence type of scoring. In this sequence-based method, the Gonnet mutation matrix can be used to optimize gap penalties, as described by Vogt and Argos (Vogt *et al.*, 1995). The second method can use a sequence-structure scoring method based on the pseudo-energy from the probe sequence "mounted" in the structural environment in the template structure. The pseudo-energy term reflects the statistical propensity of successive amino acid pairs (from the probe sequence) to be found in particular secondary structures within the template structure. The third scoring method can concern structure-structure comparisons, whereby information from the known template structure(s) is(are) compared to the predicted secondary structure of the probe sequence. A particularly preferred secondary structure prediction scheme uses a nearest neighbor algorithm.

After computing scores for the sequence-to-structure alignments, the statistical significance of the each score is preferably determined by fitting the distribution of scores to an extreme value distribution, and the raw score is compared to the chance of obtaining the same score when comparing two unrelated sequences (Jaroszewski *et al.*, 1997).

Once the alignment of the probe sequence-to-template structure has been determined, it can be used in accordance with a side chain modeling algorithm according to the invention. When a threading algorithm is used in the practice some embodiments of this invention, the probe amino acid can be "threaded" through a large database of proteins whose structures have been experimentally elucidated by, for example, x-ray crystallography or NMR spectroscopy. U.S. Patent No.

5,436,850, describes threading algorithms that can be used in the practice of this invention.

SICHO

SICHO is a new lattice protein model that represents a significant advance in our ability to computationally derive three-dimensional protein structures. In particular, SICHO focuses explicitly on the side chain center of mass positions of the amino acid residues of a target protein. The force field used in SICHO comprises short-range interactions that reflect secondary structure propensities and short-range packing biases, a geometrically implicit model of cooperative hydrogen bonds, and explicit burial, pair, and multibody tertiary interactions. When this new model force filed is combined with a small number of long-range harmonic constraints (*e.g.*, known side chain contacts), accurate three-dimensional reduced models of least medium resolution can be rapidly and efficiently generated for a given target protein.

Protein Representation

In SICHO, a target protein is modeled as a lattice chain connecting points restricted to an underlying simple cubic lattice whose mesh size equals 1.45 Å. By way of illustration, Figure 1 depicts short fragments of a β-strand and an α-helix in this particular lattice representation. This figure also shows the corresponding Cα-traces, which are not explicitly modeled by SICHO, but can be back-filled after the three-dimensional model is generated, if desired, as other or even greater levels of detail can be. The distance between two consecutive side chain units is variable and is assumed to be in the range of $11^{1/2}$-$30^{1/2}$ lattice units, or equivalently 4.8-7.9 Å. The length distribution roughly covers typical distances between two consecutive side chain centers of masse seen in real proteins.[14] The resulting number of side chain vectors, {v}, is equal to 592. Similar limitations are superimposed on the distances between the i-th and i + 2nd side chain center of mass , i-th and i + 3rd side

chain center of mass, *etc.*, up to and including the i + 8th side chain center of mass. As a result, implicit limitations are superimposed onto the range of planar angles defined by the positions of three consecutive side chains. Some possible three-vector local conformations are shown in Figure 2A.

As shown in Figure 2B, the excluded volume cluster defined for each side chain consists of the central lattice point coinciding with the hypothetical center of mass of the side chain and the 16 surrounding points located at positions (=1,0,0) and (=1,=1,0), including all permutations of these vectors. With such a hard-core definition, the distance of closest approach of two residues is equal to three lattice units (4.35 Å). This corresponds to the equivalent hard core in observed in proteins for which a high resolution three-dimensional structure has been experimentally determined. There are also 30 possible lattice positions at which the closest approach, side chain-side chain contact, can occur. These are defined by six vectors of the (3,0,0) type and 24 vectors of the (2,2,1) type emanating from the side chain of interest. For larger residues , tryptophan, phenylalanine, tyrosine, histidine, and modified side chains of similar size (with similar criteria imposed for modified side chains based in their radius of gyration), a wider, finite magnitude, repulsive core is also included, and the number of "contact positions" is even larger. Consequently, effects of lattice anisotropy are essentially nonexistent.

Side chain overlaps and interactions are readily detected by inspection of the occupancy status of the appropriate collection of lattice points in the Monte Carlo working box. As a result, for a given amino acid residue, the computational cost for calculating the short- and long-range interactions does not depend on chain length.

Monte Carlo Model and Conformational Updating

The Monte Carlo move set consists of single residue "kink" moves, chain-end moves, two-residue moves and small "rigid-body" displacements of a larger portion of the model chain. Examples of these moves are schematically illustrated in Figure 3A-D. A single "time-step" consists of N attempts at kink moves, 2